Chapter 1

Presentation attack detection in voice biometrics

Pavel Korshunov and Sébastien Marcel
Idiap Research Institute, Martigny, Switzerland
{pavel.korshunov,sebastien.marcel}@idiap.ch

Recent years have shown an increase in both the accuracy of biometric systems and their practical use. The application of biometrics is becoming widespread with fingerprint sensors in smartphones, automatic face recognition in social networks and video-based applications, and speaker recognition in phone banking and other phone-based services. The popularization of the biometric systems, however, exposed their major flaw — high vulnerability to spoofing attacks [1]. A fingerprint sensor can be easily tricked with a simple glue-made mold, a face recognition system can be accessed using a printed photo, and a speaker recognition system can be spoofed with a replay of pre-recorded voice. The ease with which a biometric system can be spoofed demonstrates the importance of developing efficient anti-spoofing systems that can detect both known (conceivable now) and unknown (possible in the future) spoofing attacks.

Therefore, it is important to develop mechanisms that can detect such attacks, and it is equally important for these mechanisms to be seamlessly integrated into existing biometric systems for practical and attack-resistant solutions. To be practical, however, an attack detection should have (i) high accuracy, (ii) be well-generalized for different attacks, and (iii) be simple and efficient.

One reason for the increasing demand for effective presentation attack detection (PAD) systems is the ease of access to people's biometric data. So often, a potential attacker can almost effortlessly obtain necessary biometric samples from social networks, including facial images, audio and video recordings, and even extract fingerprints from high resolution images. Therefore, various privacy protection solutions, such as legal privacy requirements and algorithms for obfuscating personal information, e.g., visual privacy filters [2], as well as, social awareness of threats to privacy can also increase security of personal information and potentially reduce the vulnerability of biometric systems.

In this chapter, however, we focus on presentation attacks detection in voice biometrics, i.e., automatic speaker verification (ASV) systems. We discuss vulnerabilities of these systems to presentation attacks (PAs), present different state of the art

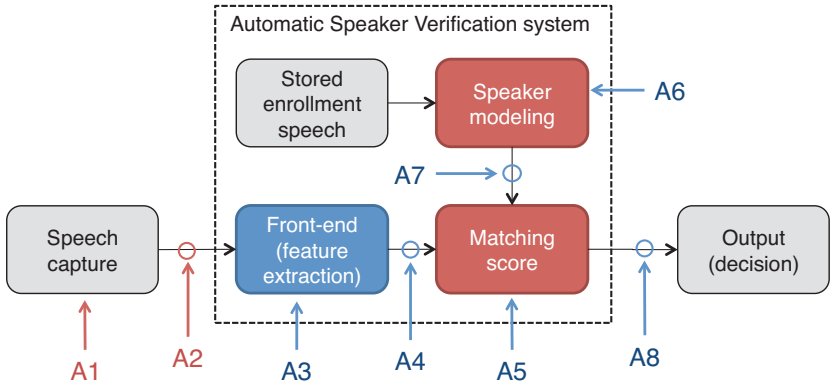


Figure 1.1: Possible attacks places in a typical ASV system.

PAD systems, give the insights into their performances, and discuss the integration of PAD and ASV systems.

1.1 Introduction

Given the complexity of a practical speaker verification system, several different modules of the system are prone to attacks, as it is identified in ISO/IEC 30107-1 standard [3] and illustrated by Figure 1.1. Depending on the usage scenario, two of the most vulnerable places for spoofing attacks in an ASV system are marked by A1 (aka 'physical access' as defined in [4] or presentation attacks) and A2 (aka 'logical access' attacks as defined in [4]) in the figure. In this chapter, we are considering A1 and A2 attacks, where the system can be attacked by presenting a spoofed signal as input. For the other points of attacks from A3 to A9, the attacker needs to have privileged access rights and know the operational details of the biometric system. Prevention of or countering such attacks is more related to system-security, and is thus out of the scope of this chapter.

There are three prominent methods through which A1 and A2 attacks can be carried out: (a) by recording and replaying the target speakers speech, (b) synthesizing speech that carries target speaker characteristics, and (c) by applying voice conversion methods to convert impostor speech into target speaker speech. Among these three, replay attack is the most viable attack, as the attacker mainly needs a recording and playback device. In the literature, it has been found that ASV systems, while immune to 'zero-effort' impostor claims and mimicry attacks [5], are vulnerable to such presentation attacks (PAs) [6]. One of the reasons for such vulnerability is a built-in ability of biometric systems in general, and ASV systems in particular, to handle undesirable variabilities. Since spoofed speech can exhibit the undesirable variabilities that ASV systems are robust to, the attacks can pass undetected.

Therefore, developing mechanisms for detection of presentation attacks is gaining interest in the speech community [7]. In that regard, the emphasis until now has

been on logical access attacks, largely thanks to the "Automatic Speaker Verification Spoofing and Countermeasures Challenge" [4], which provided a large benchmark corpus containing voice conversion-based and speech synthesis-based attacks. In the literature, development of PAD systems has largely focused on investigating short-term speech processing based features that can aid in discriminating genuine speech from spoofed signal. This includes cepstral-based features [8], phase information [9], and fundamental frequency based information, to name a few.

However, having presentation attack detection methods is not enough for practical use. Such PAD systems should be seamlessly and effectively integrated with existing ASV systems. In this chapter, we integrate speaker verification and presentation attack detection systems by using score fusion considering parallel fusion (see Figure 1.5) and cascading fusion (see Figure 1.6) schemes. The score fusion-based systems integration allows to separate *bona fide* data of the valid users, who are trying to be verified by the system, from both presentation attacks and genuine data of the non-valid users or so-called *zero-impostors*. For ASV system, we adopt verification approaches based on inter-session variability (ISV) modeling [10] and *i-vectors* [11], as the state of the art systems for speaker verification.

1.1.1 Databases

Appropriate databases are necessary for testing different presentation attack detection approaches. These databases need to contain a set of practically feasible presentation attacks and also data for speaker verification task, so that a verification system can be tested for both issues: the accuracy of speaker verification and the resistance to the attacks.

Currently, two comprehensive publicly available databases exist that can be used for vulnerability analysis of ASV systems, the evaluation of PAD methods, and evaluation of joint ASV-PAD systems: ASVspoof[1] and AVspoof[2]. Both databases contain logical access attacks (LAs), while AVspoof also contains presentation attacks (PAs). For the ease of comparison with ASVspoof, the set of attacks in AVspoof is split into LA and PA subsets (see Table 1.1).

ASVspoof database

The ASVspoof[1] database contains genuine and spoofed samples from 45 male and 61 female speakers. This database contains only speech synthesis and voice conversion attacks produced via logical access, i.e., they are directly injected in the system. The attacks in this database were generated with 10 different speech synthesis and voice conversion algorithms. Only 5 types of attacks are in the training and development set ($S1$ to $S5$), while 10 types are in the evaluation set ($S1$ to $S10$). Since last five attacks appear in the evaluation set only and PAD systems are not trained on them, they are considered 'unknown' attacks (see Table 1.1). This split of attacks allows to evaluate the systems on known and unknown attacks. The full description of

[1] http://dx.doi.org/10.7488/ds/298

[2] https://www.idiap.ch/dataset/avspoof

Table 1.1: Number of utterances in different subsets of AVspoof and ASVspoof databases.

Database	Type of data	Train	Dev	Eval
AVspoof	enroll data	780	780	868
	impostors	54509	54925	70620
	real data	4973	4995	5576
	LA attacks	17890	17890	20060
	PA attacks	38580	38580	43320
ASVspoof	enroll data	-	175	230
	impostors	-	9975	18400
	real data	3750	3497	9404
	Known attacks	12625	49875	92000
	Unknown attacks	-	-	92000

Figure 1.2: AVspoof database recording setup.

the database and the evaluation protocol are given in [4]. This database was used for the ASVspoof 2015 Challenge and is a good basis for system comparison as several systems have already been tested on it.

AVspoof database

To our knowledge, the largest publicly available database containing speech presentation attacks is AVspoof [6][2].

AVspoof database contains real (genuine) speech samples from 44 participants (31 males and 13 females) recorded over the period of two months in four sessions, each scheduled several days apart in different setups and environmental conditions such as background noises. The recording devices, including microphone AT2020USB+, Samsung Galaxy S4 phone, and iPhone 3GS, and the environments are shown in Figure 1.2. The first session was recorded in the most controlled conditions.

From the recorded genuine data, two major types of attacks were created for AVspoof database: logical access attacks, similar to those in ASVspoof database [4],

and presentation attacks. Logical access attacks are generated using (i) a statistical parametric-based speech synthesis algorithm [12] and (ii) a voice conversion algorithm from Festvox[3].

When generating presentation attacks, the assumption is that a verification system is installed on a laptop (with an internal built-in microphone) and an attacker is trying to gain access to this system by playing back to it a pre-recorded genuine data or an automatically generated synthetic data using some playback device. In AVspoof database, presentation attacks consist of (i) direct replay attacks when a genuine data is played back using a laptop with internal speakers, a laptop with external high quality speakers, Samsung Galaxy S4 phone, and iPhone 3G, (ii) synthesized speech replayed with a laptop, and (iii) converted voice attacks replayed with a laptop.

The data in AVspoof database is split into three non-overlapping subsets: training or *Train* (real and spoofed samples from 4 female and 10 male participants), development or *Dev* (real and spoofed samples from 4 female and 10 male participants), and evaluation or *Eval* (real and spoofed samples from 5 female and 11 male participants). For more details on AVspoof database, please refer to [6].

1.1.2 Evaluation

Typically, in a single database evaluation, the training subset of a given database is used for training a PAD or an ASV system. The development set is used for determining hyper-parameters of the system and evaluation set is used to test the system. In a cross-database evaluation, the training and development sets are taken from one database, while evaluation set is taken from another database. For PAD systems, a cross-attack evaluation is also possible, when the training and development sets contain one type of attack, e.g., logical access attacks only, while evaluation set contains another type, e.g., presentation or replay attacks only.

Recent recommendations 30107-3 [13] from ISO/IEC committee specify the evaluation procedure and metrics for ASV, PAD, and joint ASV-PAD systems, which we briefly present in this chapter.

ASV and joint ASV-PAD systems are evaluated under two operational scenarios: *bona fide* scenario with no attacks and the goal to separate genuine samples from zero-effort impostors and *spoof* scenario with the goal to separate genuine samples from attacks. For bona fide scenario, we report false match rate (FMR), which is similar to FAR, and false non-match rate (FNMR), which is similar to FRR, while for spoof scenario, we report impostor attack presentation match rate (IAPMR), which is the proportion of attacks that incorrectly accepted as genuine samples by the joint ASV-PAD system (for details, see recommendations in ISO/IEC 30107-3 [13]).

For evaluation of PAD systems, the following metrics are recommended: attack presentation classification error rate (APCER) and bona fide presentation classifica-

[3] http://festvox.org/

tion error rate (BPCER). APCER is the number of attacks misclassified as bona fide samples divided by the total number of attacks, and is defined as follows:

$$\text{APCER} = \frac{1}{N}\sum_{i=1}^{N}(1 - Res_i), \tag{1.1}$$

where N represents the number of attack presentations. Res_i takes value 1 if the i-th presentation is classified as an attack presentation, and value 0 if classified as a bona fide presentation. Thus, APCER can we considered as the equivalent to FAR for PAD systems, as it reflects the observed ratio of falsely accepted attack samples in relation to the total number of presented attacks.

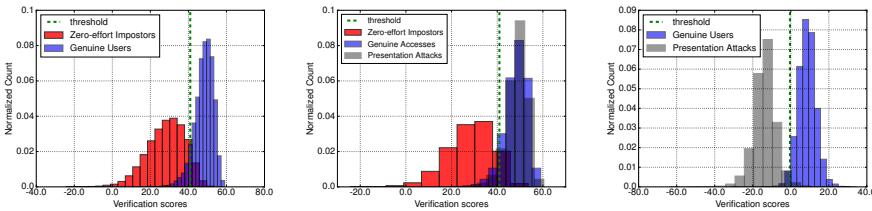
By definition, BPCER is the number of incorrectly classified bona fide (genuine) samples divided by the total number of bona fide samples:

$$\text{BPCER} = \frac{\sum_{i=1}^{N_{BF}} Res_i}{N_{BF}}, \tag{1.2}$$

where N_{BF} represents the number of bona fide presentations, and Res_i is defined similar to APCER. Thus, BPCER can we considered as the equivalent to FRR for PAD systems, as it reflects the observed ratio of falsely rejected genuine samples in relation to the total number of bona fide (genuine) samples. We compute equal error rate (EER) as the rate when APCER and BPCER are equal.

When analyzing, comparing, and especially fusing PAD and ASV systems, it is important that the scores are calibrated in a form of likelihood ratio. Raw scores can be mapped to log-likelihood ratio scores with logistic regression classifier and an associated cost of calibration C_{llr} together with a discrimination loss C_{llr}^{min} are then used as application-independent performance measures of calibrated PAD or ASV systems. Calibration cost C_{llr} can be interpreted as a scalar measure that summarizes the quality of the calibrated scores. A well-calibrated system has $0 \leq C_{llr} < 1$ and produces well-calibrated likelihood ratio. Discrimination loss C_{llr}^{min} can be viewed as the theoretically best C_{llr} value of an optimally calibrated systems. For more details on the score calibration and C_{llr} and C_{llr}^{min} metrics, please refer to [14].

Therefore, in this chapter, we report EER rates (on Eval set) when testing the considered PAD systems on each database, for the sake of consistency with the previous literature, notably [15], and BPCER and APCER of PAD systems (using the EER threshold computed on Dev set) when testing PADs in cross-database scenario. EER has been commonly used within the speech community to measure the performance of ASV and PAD systems, while BPCER and APCER are the newly standardized metrics and we advocate for the use of the open evaluation standards in the literature. We also report calibration cost C_{llr} and the discrimination loss C_{llr}^{min} metrics for the individual PAD systems. FMR, FNMR, and IAPMR are reported for ASV and joint ASV-PAD systems on evaluation set (using EER threshold computed on the development set).



(a) ASV *bona fide* scenario (b) ASV *spoof* scenario (c) PAD system

Figure 1.3: Histogram distributions of scores from *i-vector* based ASV system in *bona fide* and *spoof* scenario, and *MFCC*-based PAD system.

1.2 Vulnerability of voice biometrics

The research on automatic speaker verification (ASV) is more established with regular competitions conducted by National Institute of Standards and Technology (NIST) since 1996[4]. Many techniques have been proposed with the most notable systems based on Gaussian mixture model (GMM), inter-session variability (ISV) modeling [10], joint factor analysis (JFA) [16], and *i-vectors* [11].

To demonstrate vulnerability of ASV systems to presentation attacks, we consider two systems based on inter-session variability (ISV) modeling [10] and *i-vectors* [11], which are the state of the art speaker verification systems able to effectively deal with intra-class and inter-class variability. In these systems, voice activity detection is based on the modulation of the energy around 4Hz, the features include 20 mel-scale frequency coefficients (MFCC) and energy, with their first and second derivatives, and modeling was performed with 256 Gaussian components using 25 expectation-maximization (EM) iterations. Universal background model (UBM) was trained using training set of publicly available MOBIO database[5], while the clients models are build using an enrollment data from the development set of AVspoof database (only genuine data).

In *i-vectors* based system, for a given audio sample, the supervector of GMM mean components (computed for all frames of the sample) is reduced to an *i-vector* of the dimension 100, which essentially characterizes the sample. These *i-vectors* are compensated for channel variability using linear discriminative analysis (LDA) and within class covariance normalization (WCCN) techniques (see [11] for more details).

Table 1.2 demonstrates how *i-vectors* and ISV-based ASV systems perform in two different scenarios: (i) when there are no attacks present (zero-impostors only), referred to as *bona fide* scenario (defined by ISO/IEC [13]), and (ii) when the system is being spoofed with presentation attacks, referred to as *spoof* scenario. Histograms of score distribution in Figure 1.3b also illustrate the effect of attacks on *i-vectors* based ASV system in *spoof* scenario, compared to *bona fide* scenario in Figure 1.3a.

[4] http://www.nist.gov/itl/iad/mig/sre.cfm

[5] https://www.idiap.ch/dataset/mobio

Table 1.2: ISV-based and *i-vector* ASVs on evaluation set of AVspoof database.

ASV system	**Zero-impostors only**		**PAs only**
	FMR (%)	**FNMR (%)**	**IAPMR (%)**
ISV-based	4.46	9.90	92.41
i-vectors based	8.85	8.31	94.04

From Table 1.2, it can be noted that both ASV systems perform relatively well under *bona fide* scenario with ISV-based system showing lower FMR of 4.46%. However, when a spoofed data is introduced, without a PAD system in place, the IAPMR significantly increases reaching 92.41% for ISV-based and 94.04% for *i-vectors* based systems. It means that a typical verification system is not able to correctly distinguish presentation attacks from genuine data.

1.3 Presentation attack detection approaches

As was shown in the previous section, ASV systems are highly susceptible to presentation attacks. This vulnerability motivated researchers to propose different systems and methods for detecting such attacks (see Figure 1.4). In this section, we present the most commonly used recent approaches and discuss feature extraction and classification components, as well as, score fusion integration technique.

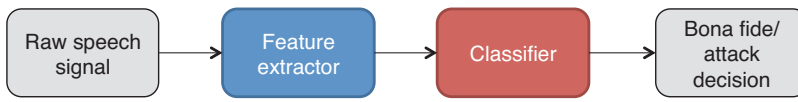


Figure 1.4: Presentation attack detection system.

1.3.1 Features

A survey by Wu *et al.* [7] provides a comprehensive overview of both the existing spoofing attacks and the available attack detection approaches. An overview of the methods for synthetic speech detection by Sahidullah *et al.* [15] benchmarks several existing feature extraction methods and classifiers on ASVspoof database.

Existing approaches to feature extraction for speech spoofing attack detection methods include spectral- and cepstral-based features [8], phase-based features [9], the combination of amplitude and phase features of the spectrogram [17], and audio quality based features [18]. Features directly extracted from a spectrogram can also be used, as per the recent work that relies on local maxima of spectrogram [19].

Compared to cepstral coefficients, using phase information extracted from the signal seem to be more effective for anti-spoofing detection, as it was shown by De Leon *et al.* [9] and Wu *et al.* [20]. However, the most popular recent approaches rely on the combination of spectral-based and phase-based features [17, 21, 22, 23].

Most of these features are used successfully in speaker verification systems already, so, naturally, they are first to be proposed for anti-spoofing systems as well.

In addition to these spectral-based features, features based on pitch frequency patterns have been proposed [24, 25]. There are also methods that aim to extract "pop-noise" related information that is indicative of the breathing effect inherent in normal human speech [8].

Constant Q cepstral coefficients (CQCCs) [26] features were proposed recently and they have shown a superior performance in detecting both know and unknown attacks in ASVspoof database. Also, a higher computational layer can be added, for instance, Alegre *et al.* [27] proposed to use histograms of Local Binary Patterns (LBP), which can be computed directly from a set of pre-selected spectral, phase-based, or other features.

1.3.2 Classifiers

Besides determining 'good features for detecting presentation attacks', it is also important to correctly classify the computed feature vectors as belonging to bona fide or spoofed data. Choosing a reliable classifier is especially important given a possibly unpredictable nature of attacks in a practical system, since it is unknown what kind of attack the perpetrator may use when spoofing the verification system. The most common approach to classification is to use one of the well-known classifiers, which is usually pre-trained on the examples of both real and spoofed data. To simulate realistic environments, the classifier can be trained on a subset of the attacks, termed *known attacks*, and tested on a larger set of attacks that include both known and *unknown attacks*.

Different methods use different classifiers but the most common choices include logistic regression, support vector machine (SVM), and Gaussian mixture model (GMM) classifiers. The benchmarking study on logical access attacks [15] finds GMMs to be more successful compared to two-class SVM (combined with an LBP-based feature extraction from [27]) in detecting synthetic spoofing attacks. Deep learning networks are also showing promising performance in simultaneous feature selection and classification [28].

1.3.3 Fusion

Fusion of different features or the results of different classification systems is a natural way of combining different systems, in our case, PAD and ASV systems to create a joint verification system resistant to the attacks.

In this chapter, we focus on a score level fusion as a means to integrate different ASV and PAD systems into one joint system. Due to relative simplicity of such fusion and the evidence that it leads to a better performing combined systems, this operation has become popular among researchers. However, the danger is to rely on score fusion blindly without studying how it can affect different systems in different scenarios.

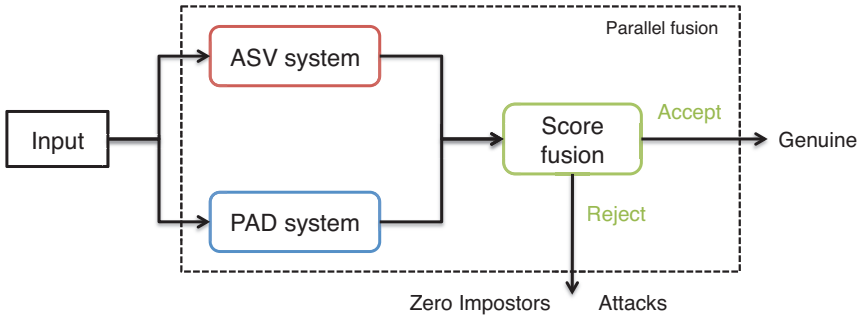


Figure 1.5: A joint ASV-PAD system based on parallel score fusion.

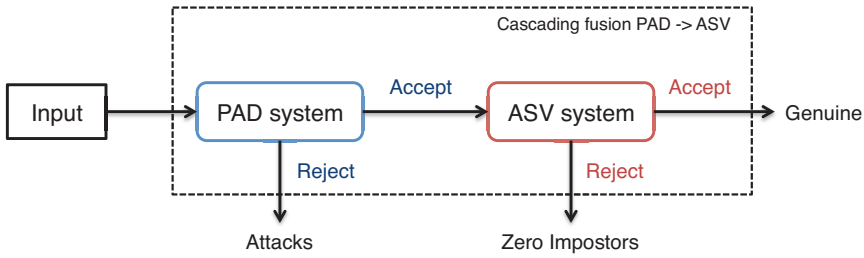


Figure 1.6: A joint PAD-ASV system based on cascading score fusion (reversed order of the systems leads to the same results).

One way to fuse ASV and PAD systems at the score level is to use a parallel scheme, as it is illustrated in Figure 1.5. In this case, the scores from each of N system are combined into a new feature vector of length N that need to be classified. The classification task can be performed using different approaches, and, in this chapter, we consider three different algorithms: (i) a logistic regression classifier, denoted as 'LR', which leads to a straight line separation, as illustrated by the scatter plot in Figure 1.7a, (ii) a polynomial logistic regression, denoted as 'PLR', which results in a polynomial separation line, and (iii) a simple mean function, denoted as 'Mean', which is taken on scores of the fused systems. For 'LR' and 'PLR' fusion, the classifier is pre-trained on the score-feature vectors from a training set.

Another common way to combine PAD and ASV systems is a cascading scheme, in which one system is used first and only the samples that are accepted by this system (based on its own threshold) are then passed to the second system, which will further filter the samples, using its own independently determined threshold. Effectively, cascading scheme can be viewed as a *logical and* of two independent systems. Strictly speaking, when considering one PAD and one ASV systems, there are two variants of cascading scheme: (i) when ASV is used first, followed by PAD, and (ii) when PAD is used first, followed by ASV (see Figure 1.6). Although these schemes are equivalent, i.e., *and* operation is commutative, and they both lead to

the same filtering results (the same error rates), we consider variant (ii), since it is defined in ISO/IEC 30107-1 standard [3].

When using a score level fusion, it is important to perform a thorough evaluation of the combined/fused system to understand how incorporating PAD system affects verification accuracy for both real and spoofed data. In the upcoming parts of this chapter, we therefore adopt and experimentally apply an evaluation methodology specifically designed for performance assessment of fusion system proposed in [29].

1.4 PADs failing to generalize

To demonstrate the performance of PAD systems in single database and in cross-database scenario, we have selected several state of the art methods for presentation attacks detection in speech, which were recently evaluated by Sahidullah *et al.* [15] on ASVspoof database with an addition of CQCC features based method [26].

These systems rely on GMM-based classifier (two models for real and attacks, 512 Gaussians components with 10 EM iterations for each model), since it has demonstrated improved performance compared to support vector machine (SVM) on the data from ASVspoof database. Four cepstral-based features with mel-scale, i.e.,, mel-frequency cepstral coefficients (MFCC) [30], rectangular (RFCC), inverted mel-scale (IMFCC), and linear (LFCC) filters [31] were selected. These features are computed from a power spectrum (power of magnitude of 512-sized fast Fourier transform) by applying one of the above filters of a given size (we use size 20 as per [15]). Spectral flux-based features, i.e., subband spectral flux coefficients (SSFC) [32], which are Euclidean distances between power spectrums (normalized by the maximum value) of two consecutive frames, subband centroid frequency (SCFC) [33], and subband centroid magnitude (SCMC) [33] coefficients are considered as well. A discrete cosine transform (DCT-II) is applied to these above features, except for SCFC, and first 20 coefficients are taken. Before computing selected features, a given audio sample is first split into overlapping 20ms-long speech frames with 10ms overlap. The frames are pre-emphasized with 0.97 coefficient and pre-processed by applying Hamming window. Then, for all features, deltas and double-deltas [34] are computed and only these derivatives (40 in total) are used by the classifier. Only deltas and deta-deltas are kept, because [15] reported that static features degraded performance of PAD systems.

In addition to the above features, we also consider recently proposed CQCC [26], which are computed using constant Q transform instead of FFT. To be consistent with the other features and fair in the systems comparison, we used also only delta and delta-deltas (40 features in total) derived from 19 plus C_0 coefficients.

The selected PAD systems are evaluated on each ASVspoof and AVspoof database and in cross-database scenario. To keep results comparable with current state of the art work [15, 35], we computed average EER (Eval set) for single database evaluations and APCER with BPCER for cross-database evaluations. APCER with BPCER are computed for Eval set of a given dataset using the EER threshold obtained from the Dev set from another dataset (see Table 1.4).

Table 1.3: Performance of PAD systems in terms of average EER (%), C_{llr}, and C_{llr}^{min} of calibrated scores for evaluation sets of ASVspoof [4] and AVspoof [6] databases.

PADs	**ASVspoof (Eval)**							**AVspoof (Eval)**					
	Known			**S10**	**Unknown**			**LA**			**PA**		
	EER	C_{llr}	C_{llr}^{min}	EER	EER	C_{llr}	C_{llr}^{min}	EER	C_{llr}	C_{llr}^{min}	EER	C_{llr}	C_{llr}^{min}
SCFC	0.11	0.732	0.006	23.92	5.17	0.951	0.625	0.00	0.730	0.000	5.34	0.761	0.160
RFCC	0.14	0.731	0.009	6.34	1.32	0.825	0.230	0.04	0.729	0.001	3.27	0.785	0.117
LFCC	0.13	0.730	0.005	5.56	1.20	0.818	0.211	0.00	0.728	0.000	4.73	0.811	0.153
MFCC	**0.47**	**0.737**	**0.023**	**14.03**	**2.93**	**0.877**	**0.435**	**0.00**	**0.727**	**0.000**	**5.43**	**0.812**	**0.165**
IMFCC	0.20	0.730	0.007	5.11	1.57	0.804	0.192	0.00	0.728	0.000	4.09	0.797	0.137
SSFC	0.27	0.733	0.016	7.15	1.60	0.819	0.251	0.70	0.734	0.027	4.70	0.800	0.160
SCMC	0.19	0.731	0.009	6.32	1.37	0.812	0.229	0.01	0.728	0.000	3.95	0.805	0.141
CQCC	0.10	0.732	0.008	1.59	0.58	0.756	0.061	0.66	0.733	0.028	3.84	0.796	0.128

Table 1.4: Performance of PAD systems in terms of average APCER (%), BPCER (%), and C_{llr} of calibrated scores in cross-database testing on ASVspoof [4] and AVspoof [6] databases.

PADs	**ASVspoof (Train/Dev)**						**AVspoof-LA (Train/Dev)**					
	AVspoof-LA (Eval)			**AVspoof-PA (Eval)**			**ASVspoof (Eval)**			**AVspoof-PA (Eval)**		
	APCER	BPCER	C_{llr}	APCER	BPCER	C_{llr}	APCER	BPCER	C_{llr}	APCER	BPCER	C_{llr}
SCFC	0.10	2.76	0.751	10.20	2.76	0.809	15.12	0.00	0.887	39.62	0.35	0.970
RFCC	0.29	69.57	0.887	7.51	69.57	0.927	26.39	0.00	0.902	48.32	2.86	0.988
LFCC	1.30	0.13	0.740	21.03	0.13	0.868	17.70	0.00	0.930	37.49	0.02	0.958
MFCC	**1.20**	**2.55**	**0.764**	**17.09**	**2.55**	**0.838**	**10.60**	**0.00**	**0.819**	**19.72**	**1.22**	**0.870**
IMFCC	4.57	0.00	0.761	92.98	0.00	1.122	99.14	0.00	1.164	43.00	0.60	0.966
SSFC	4.81	64.47	0.899	18.89	64.47	0.973	71.84	0.68	1.047	63.45	23.54	1.070
SCMC	0.75	1.70	0.750	22.61	1.70	0.866	15.94	0.00	0.861	45.97	0.01	0.978
CQCC	13.99	57.05	0.968	66.29	57.05	1.191	44.65	0.61	1.009	0.86	100.00	1.009

To avoid prior to the evaluations, the raw scores from each individual PAD system are pre-calibrated with logistic regression based on Platts sigmoid method [36] by modeling scores of the training set and applying the model on the scores from development and evaluation sets. The calibration cost C_{llr} and the discrimination loss C_{llr}^{min} of the resulted calibrated scores are provided.

In Table 1.3, the results for known and unknown attacks (see Table 1.1) of *Eval* set of ASVspoof are presented separately to demonstrate the differences between these two types of attacks provided in ASVspoof database. The main contribution to the higher EER of unknown is given by a more challenging attack 'S10' of the evaluation set.

Since AVspoof contains both logical access (LA for short) and presentation attacks (PA), the results for these two types of attacks are also presented separately. Hence, it allows to compare the performance on ASVspoof database (it has logical access attacks only) with an AVspoof-LA attacks.

From the results in Table 1.3, we can note that (i) LA set of AVspoof is less challenging compared to ASVspoof for almost all methods, (ii) unknown attacks for which PAD is not trained is more challenging, and (iii) presentation attacks are also more challenging compared to LA attacks.

Table 1.4 presents the cross-database results when a given PAD system is trained and tuned using training and development sets from one database but is tested using evaluation set from another database. For instance, results in the second column of the table are obtained by using training and development sets from ASVspoof database but evaluation set from AVspoof-LA. Also, we evaluated the effect of using one type of attacks (e.g., logical access from AVspoof-LA) for training and another type (e.g., presentation attacks of AVspoof-PA) for testing (the results are in the last column of the table).

From the results in Table 1.4, we can note that all methods generalize poorly across different datasets with BPCER reaching 100%, for example, CQCC-based PAD shows poor performance for all cross-database evaluations. It is also interesting to note that even similar methods, for instance, RFCC and MFCC-based, have very different accuracy in cross-database testing, even though they showed less drastic difference in single-database evaluations (see Table 1.3).

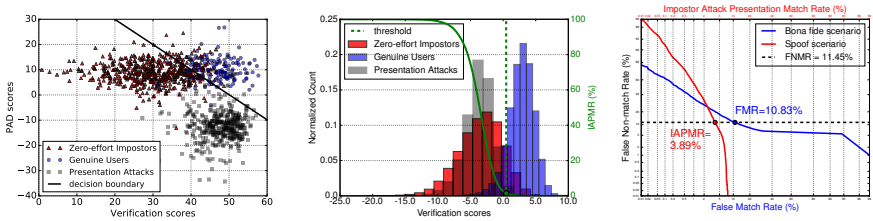
1.5 Integration of PAD and ASV

As described in Section 1.3, multiple presentation attack detection systems have been considered to detect whether a given speech sample is real or spoofed. However, the purpose of a PAD system is to work in tandem with a verification system, so that the joint system can effectively separate the genuine data from both zero-effort impostors (genuine data but incorrect identity) and spoofed attacks (spoofed data for the correct identity).

Table 1.5: Fusing *i-vector* and ISV-based verification systems with the selected MFCC-based PAD (in bold in Tables 1.3 and 1.4) on evaluation set of AVspoof-PA.

ASV system	**Fused with PAD**	**Type of fusion**	**Zero-impostors only**		**PAs only**
			FMR (%)	**FNMR (%)**	**IAPMR (%)**
ISV-based	no fusion	-	4.46	9.90	92.41
	MFCC	**Cascade**	**6.57**	**12.00**	**4.19**
	MFCC	Mean	23.05	22.73	28.98
	MFCC	LR	25.40	24.72	2.68
	MFCC	PLR	4.97	10.75	5.17
midrule *i-vectors* based	no fusion	-	8.85	8.31	94.04
	MFCC	**Cascade**	**10.83**	**11.45**	**3.89**
	MFCC	Mean	26.33	19.44	19.47
	MFCC	LR	8.77	8.33	94.28
	MFCC	PLR	9.60	10.47	95.76

As presented in Section 1.3.3, in a score-based fusion of PAD and ASV systems, we make a decision about each speech sample using the scores from both PAD and ASV. The resulted joint system can effectively distinguish genuine data from presentation attacks, as demonstrated in Figure 1.7b for ASV based on *i-vector* integrated with an example of MFCC-based PAD system. We have chosen MFFC-based system as an example for Figure 1.7, because, from the Table 1.5 it is clear that applying



(a) Scatter plot, Test set (b) Histogram distribution, Test set (c) DET curves, Test set

Figure 1.7: A scatter plot, histogram distributions, and DET curves for joint *i-vector* ASV and MFCC-based PAD systems.

cascade fusion scheme to join an ASV system with *MFCC* leads to more superior performance compared to other fusion schemes and algorithms.

As results presented in Table 1.5 demonstrate, integration with PAD system can effectively reduce IAPMR from above 90% of the ASV (both ISV-based and *i-vector*) down to 3.89%, which is the best performing system of *i-vector* ASV fused with *MFCC*-based PAD via cascade fusion (see Figure 1.7c for DET plots of different scenarios). Such drastic improvement in the attack detection comes with an increase in FMR (from 4.46% to 6.57% when ASV is ISV and from 8.85% to 10.83% when ASV is *i-vector*). FNMR also increases.

Please note that an important advantage of using MFCC-based PAD is that MFCC are the most commonly used fast to compute features in speech processing, which makes it practical to use MFCC-based PAD for fusion with an ASV.

The Table 1.5 also shows that cascading fusion leads to a better overall performance compared to parallel scheme. However, compared to a cascading scheme, where each fused system is independent and has to be tuned separately for disjoint set of parameter requirements, parallel scheme is more flexible, because it allows to tune several parameters of the fusion, as if it was one single system consisting of interdependent components. Such flexibility can be valuable in practical systems. See [29] for a detailed comparison of the different fusion schemes and their discussion.

1.6 Conclusions

In this chapter, we provide an overview of the existing presentation attack detection (PAD) systems for voice biometrics and present evaluation results for selected eight systems on two most comprehensive publicly available databases, AVspoof and ASVspoof. The cross-database evaluation results of these selected methods demonstrate that state of the art PAD systems generalize poorly across different databases and data. The methods generalize especially poorly, when they were trained on 'logical access' attacks and tested on more realistic presentation attacks, which means a new and more practically applicable attack detection methods need to be developed.

We also consider score-based integration of several PAD and ASV systems following both cascading and parallel schemes. Presented evaluation results show a significantly increased resistance of joined ASV-PAD systems to presentation attacks from AVspoof database, with cascading fusion leading to a better overall performance compared to parallel scheme.

Presentation attack detection in voice biometrics is far from being solved, as currently proposed methods do not generalize well across different data. It means that no effective method is yet proposed that would make speaker verification system resistant even to trivial replay attacks, which prevents the wide adoption of ASV systems in practical applications, especially in security sensitive areas. Deep learning methods for PAD are showing some promise and may be able to solve the issue of generalizing across different attacks.

Acknowledgements

This work was conducted in the framework of EU H2020 project TeSLA, Norwegian SWAN project, and Swiss Centre for Biometrics Research and Testing.

Bibliography

[1] S. Marcel, M. S. Nixon, and S. Z. Li, *Handbook of Biometric Anti-Spoofing: Trusted Biometrics Under Spoofing Attacks*. Springer Publishing Company, Incorporated, 2014.

[2] P. Korshunov and T. Ebrahimi, "Towards optimal distortion-based visual privacy filters," in *IEEE International Conference on Image Processing*, 2014.

[3] ISO/IEC JTC 1/SC 37 Biometrics, "DIS 30107-1, information technology – biometrics presentation attack detection," American National Standards Institute, Jan. 2016.

[4] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 2037–2041.

[5] J. Mariéthoz and S. Bengio, "Can a professional imitator fool a GMM-based speaker verification system?" IDIAP, Tech. Rep. Idiap-RR-61-2005, 2005.

[6] S. Kucur Ergunay, E. Khoury, A. Lazaridis, and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," in *IEEE International Conference on Biometrics: Theory, Applications and Systems*, Sep. 2015.

[7] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[8] S. Shiota, F. Villavicencio, J. Yamagishi, N. Ono, I. Echizen, and T. Matsui, "Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[9] P. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of hmm-based synthetic speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 8, pp. 2280–2290, Oct 2012.

[10] R. Vogt and S. Sridharan, "Explicit modelling of session variability for speaker verification," *Comput. Speech Lang.*, vol. 22, no. 1, pp. 17–38, Jan. 2008. [Online]. Available: http://dx.doi.org/10.1016/j.csl.2007.05.003

[11] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[12] H. Zen, K. Tokuda, and A. W. Black, "Review: Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[13] ISO/IEC JTC 1/SC 37 Biometrics, "DIS 30107-3:2016, information technology – biometrics presentation attack detection — part 3: Testing and reporting," American National Standards Institute, Oct. 2016.

[14] M. I. Mandasari, M. Gnther, R. Wallace, R. Saeidi, S. Marcel, and D. A. van Leeuwen, "Score calibration in face recognition," *IET Biometrics*, vol. 3, no. 4, pp. 246–256, 2014.

[15] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. of Interspeech*, 2015.

[16] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, May 2007.

[17] T. B. Patel and H. A. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech," in *INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 2062–2066.

[18] A. Janicki, "Spoofing countermeasure based on analysis of linear prediction error," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[19] J. Gaka, M. Grzywacz, and R. Samborski, "Playback attack detection for text-dependent speaker verification over telephone channels," *Speech Communication*, vol. 67, pp. 143 – 153, 2015.

[20] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 7234–7238.

[21] M. J. Alam, P. Kenny, G. Bhattacharya, and T. Stafylakis, "Development of crim system for the automatic speaker verification spoofing and countermeasures challenge 2015," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[22] L. Wang, Y. Yoshida, Y. Kawakami, and S. Nakagawa, "Relative phase information for detecting human speech and spoofed speech," in *Proc. of Interspeech 2015*, 2015.

[23] Y. Liu, Y. Tian, L. He, J. Liu, and M. T. Johnson, "Simultaneous utilization of spectral magnitude and phase information to extract supervectors for speaker verification anti-spoofing," *Proc. of Interspeech 2015*, vol. 2, p. 1, 2015.

[24] P. L. De Leon, B. Stewart, and J. Yamagishi, "Synthetic speech discrimination using pitch pattern statistics derived from image analysis." in *Proc. of Interspeech*, 2012, pp. 370–373.

[25] A. Ogihara, U. Hitoshi, and A. Shiozaki, "Discrimination method of synthetic speech using pitch frequency against synthetic speech falsification," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 88, no. 1, pp. 280–286, 2005.

[26] M. Todisco, H. Delgado, and N. Evans, "Articulation rate filtering of CQCC features for automatic speaker verification," in *INTERSPEECH*, San Francisco, USA, 09 2016.

[27] F. Alegre, A. Amehraye, and N. Evans, "A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns," in *Proc. of BTAS*, Sept 2013, pp. 1–8.

[28] D. Luo, H. Wu, and J. Huang, "Audio recapture detection using deep learning," in *Signal and Information Processing (ChinaSIP), 2015 IEEE China Summit and International Conference on*, July 2015, pp. 478–482.

[29] I. Chingovska, A. Anjos, and S. Marcel, "Biometrics evaluation under spoofing attacks," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2264–2276, Dec 2014.

[30] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug 1980.

[31] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 254–272, Apr 1981.

[32] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. of ICASSP*, vol. 2, Apr 1997, pp. 1331–1334 vol.2.

[33] P. N. Le, E. Ambikairajah, J. Epps, V. Sethu, and E. H. C. Choi, "Investigation of spectral centroid features for cognitive load classification," *Speech Communication*, vol. 53, no. 4, pp. 540–551, Apr. 2011.

[34] F. K. Soong and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 6, pp. 871–879, Jun 1988.

[35] U. Scherhag, A. Nautsch, C. Rathgeb, and C. Busch, "Unit-selection attack detection based on unfiltered frequency-domain features," in *INTERSPEECH*, San Francisco, USA, 09 2016, p. 2209.

[36] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in large margin classifiers*. MIT Press, 1999, pp. 61–74.